

1 Training deep learning algorithms with weakly labeled 2 pneumonia chest X-ray data for COVID-19 detection

3 Sivaramakrishnan Rajaraman * and Sameer Antani

4 Lister Hill National Center for Biomedical Communications, National Library of Medicine, 8600 Rockville
5 Pike, Bethesda, MD 20894, USA; santani@mail.nih.gov

6 * Correspondence: sivaramakrishnan.rajaraman@nih.gov; Tel.: +1-301-827-2383

7 **Abstract:** The novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has caused
8 a pandemic resulting in over 2.7 million infected individuals and over 190,000 deaths and growing.
9 Respiratory disorders in COVID-19 caused by the virus commonly present as viral pneumonia-like
10 opacities in chest X-ray images which are used as an adjunct to the reverse transcription-polymerase
11 chain reaction test for confirmation and evaluating disease progression. The surge places high
12 demand on medical services including radiology expertise. However, there is a dearth of sufficient
13 training data for developing image-based automated decision support tools to alleviate radiological
14 burden. We address this insufficiency by expanding training data distribution through use of
15 weakly-labeled images pooled from publicly available CXR collections showing pneumonia-related
16 opacities. We use the images in a stage-wise, strategic approach and train convolutional neural
17 network-based algorithms to detect COVID-19 infections in CXRs. It is observed that weakly-
18 labeled data augmentation improves performance with the baseline test data compared to non-
19 augmented training by expanding the learned feature space to encompass variability in the unseen
20 test distribution to enhance inter-class discrimination, reduce intra-class similarity and
21 generalization error. Augmentation with COVID-19 CXRs from individual collections significantly
22 improves performance compared to baseline non-augmented training and weakly-labeled
23 augmentation toward detecting COVID-19 like viral pneumonia in the publicly available COVID-
24 19 CXR collections. This underscores the fact that COVID-19 CXRs have a distinct pattern and hence
25 distribution, unlike non-COVID-19 viral pneumonia and other infectious agents.

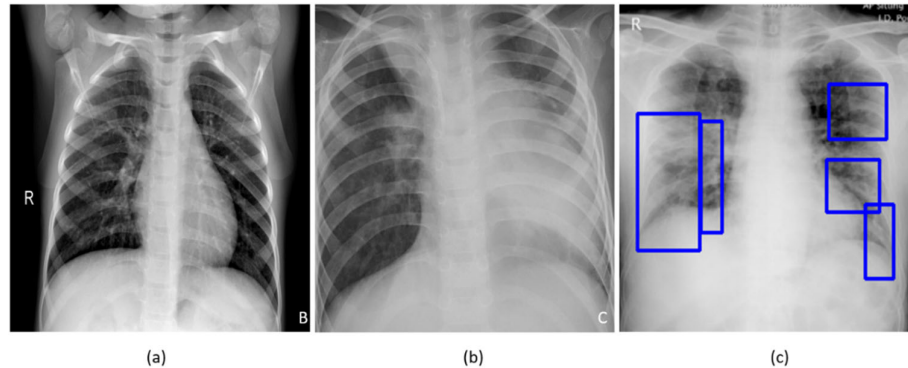
26 **Keywords:** augmentation; chest-X-rays; convolutional neural network; COVID-19; deep learning;
27 pneumonia; localization

29 1. Introduction

30 The novel Coronavirus disease 2019 (COVID-19) is caused by a strain of coronavirus called the
31 Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) that originated in Wuhan in the
32 Hubei province in China. On March 11, 2020, the World Health Organization (WHO) declared the
33 disease as a pandemic [1], and as of this writing (in late April 2020), there are more than 2.7 million
34 globally confirmed cases with over 190,000 reported deaths with unabated growth. The disease is
35 detected using the reverse transcription-polymerase chain reaction (RT-PCR) tests that are shown to
36 exhibit high specificity but variable sensitivity in detecting the presence of the disease [2]. However,
37 these test kits are in limited supply in some geographical regions, particularly third-world countries
38 [3]. The turnaround time is reported to be 24 hours in major cities and even greater in rural regions.
39 This necessitates the need to explore other options to identify the disease and facilitate swift referrals
40 for the COVID-19 affected patient population in need of urgent medical care.

41 A study of literature shows that viral pneumonia is commonly found to affect the lungs with the
42 progression of COVID-19 disease, often manifesting as ground-glass opacities (GGO), with
43 peripheral, bilateral, and predominant basal distribution in the lungs, preventing oxygen entry,
44 thereby causing breathing difficulties along with hyperthermia [2]. These patterns are visually similar
45 to, yet distinct from those caused by non-COVID-19-related viral pneumonia and those caused by
46 other bacterial and fungal pathogens [2]. Also, current literature studies reveal that it is difficult to
47 distinguish viral pneumonia from others caused by bacterial and fungal pathogens [4]. Fig. 1 shows

48 instances of CXRs with clear lungs, showing bacterial pneumonia, and COVID-19-related
49 pneumonia, respectively.
50



51
52 **Figure 1.** CXRs showing (a) Clear lungs; (b) Bacterial pneumonia infections manifesting as
53 consolidations in the right upper lobe and retro-cardiac left lower lobe; (c) COVID-19 pneumonia
54 infection showing bilateral manifestations.
55

56 While not recommended as a primary diagnostic tool due to risk of increased transmission, chest
57 radiography and computed tomography (CT) scans are used to screen/confirm respiratory damage
58 in COVID-19 disease and evaluate its progression [3]. CT scans are reported to be less specific than
59 RT-PCR but highly sensitive in detecting COVID-19 and can play a pivotal role in disease
60 diagnosis/treatment [3]. However, the American College of Radiology has recommended against use
61 of CT scans as a first-line test¹. Additional considerations of increased risk of transmission, access,
62 and cost also contribute to the recommendation. When radiological imaging is considered necessary,
63 portable chest X-rays (CXRs) are considered a good and viable alternative [2]. However, in a
64 pandemic situation, assessment of the images places a huge burden on radiological expertise, which
65 is often lacking in regions with limited resources. Automated decision-making tools could be
66 valuable in alleviating some of this burden, and also as a research tool for quantifying disease
67 progression.

68 A study of literature shows that automated computer-aided diagnostic (CADx) tools built with
69 data-driven deep learning (DL) algorithms using convolutional neural networks (CNN) have shown
70 promise in detecting, classifying, and quantifying COVID-19-related disease patterns using CXRs
71 and CT scans [5, 6] and can serve as a triage under resource-constrained settings thereby facilitating
72 swift referrals that need urgent patient care. These tools combine elements of radiology and computer
73 vision to learn the hierarchical feature representations from medical images to identify typical disease
74 manifestations and localize suspicious regions of interest (ROI).

75 It is customary to train and test a DL model with the data coming from the same target
76 distribution to offer probabilistic predictions toward categorizing the medical images to their
77 respective categories. Often this idealized target is not possible due to limited data availability, or
78 weak labels. In the present situation, despite a large number of cases worldwide, we have very limited
79 COVID-19 CXR image data that is publicly available to train DL models where the goal is to recognize
80 CXR images showing COVID-19-related viral pneumonia from those caused by other non-COVID-
81 19 viral, bacterial and other pathogens. Acquiring such data remains a goal for medical societies such
82 as the Radiological Society of North America (RSNA)² and Imaging COVID-19 AI Initiative in
83 Europe³. Large number of training data enable a diversified feature space across categories that help

¹<https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection>

²https://press.rsna.org/timssnet/media/pressreleases/14_pr_target.cfm?ID=2167

³<https://imagingcovid19ai.eu/>

84 enhance inter-class variance leading to better DL performance. The absence of such data leads to
85 model overfitting and poor generalization to unseen real-world data. Under these circumstances,
86 data augmentation has been proven to be effective in training discriminative DL models [7]. There
87 are several data augmentation methods discussed in the literature for improving performance in
88 natural computer vision tasks. These include traditional augmentation techniques like flipping,
89 rotations, color jittering, random cropping, and elastic distortions and generative adversarial
90 networks (GAN) based synthetic data generation [8].

91 Unlike natural images, such as those found in ImageNet [9], medical images tend to have
92 different visual characteristics exhibiting high inter-class similarities and highly localized ROI. Under
93 these circumstances, traditional augmentation methods that introduce simple pixel-wise image
94 modifications are shown to be less effective [10]. On the other hand, GAN-based DL models that are
95 used for synthetic data generation are computationally complex and the jury is still out on the
96 anatomical and pathological validity of synthesized images. These networks are hard to train due to
97 the problem of Nash equilibria, defined as the zero-sum game between the generator and the
98 discriminator networks where they contest with each other in improving performance [11]. Further,
99 these networks are shown to be sensitive to the selection of architecture and hyperparameters and
100 often get into mode collapse, resulting in degraded performance [11]. In general, there is a great
101 opportunity for research in developing effective data augmentation strategies for medical visual
102 recognition tasks. Goals for such medical data augmentation techniques include reducing overfitting
103 and regularization errors in a data-scarce situation. The urgency offered by the pandemic has led to
104 the motivation behind this study.

105 In this work, we use weakly-labeled CXR images that are pooled from publicly available
106 collections showing pneumonia-related opacities to augment training data toward improving inter-
107 class variance. The goal is to improve COVID-19 detection in CXRs, with the baseline being the
108 training data without augmentation.

109 **2. Materials and Methods**

110 *2.1. Data and Workflow*

111 This retrospective analysis is performed using four publicly available CXR collections:

112 A) Pediatric CXR dataset [4]: A set of 5,232 anterior-posterior (AP) projection CXR images of
113 children of 1 to 5 years of age acquired as part of routine clinical care at the Guangzhou Children's
114 Medical Center in China. The set contains 1583 normal, 2780 bacterial pneumonia, and 1493 CXRs
115 showing non-COVID-19 viral pneumonia, respectively.

116 B) RSNA CXR dataset [12]: The RSNA, Society of Thoracic Radiology (STR), and the National
117 Institutes of Health (NIH) jointly organized the Kaggle pneumonia detection challenge to develop
118 image analysis and machine learning algorithms to automatically categorize the CXRs as showing
119 normal, non-pneumonia-related or pneumonia-related opacities. The publicly available data is a
120 curated subset of 26,684 AP and posterior-anterior (PA) CXRs showing normal and abnormal
121 radiographic patterns, taken from the NIH CXR-14 dataset [13]. It includes 6012 CXRs showing
122 pneumonia-related opacities with ground truth (GT) bounding box annotations for these on 1,241
123 CXRs.

124 C) CheXpert CXR dataset [14]: A subset of 4683 CXRs showing pneumonia-related opacities
125 selected from a collection of 223,648 CXRs in frontal and lateral projections, collected from 65,240
126 patients at Stanford Hospital, California, and labeled for 14 thoracic diseases by extracting the labels
127 from radiological texts using an automated natural language processing (NLP)-based labeler,
128 conforming to the glossary of the Fleischner Society.

129 D) NIH CXR-14 dataset [13]: A subset of 307 CXRs showing pneumonia-related opacities
130 selected from a collection of 112,120 CXRs in frontal projection, collected from 30,805 patients. Images
131 are labeled with 14 thoracic disease labels extracted automatically from radiological reports using an
132 NLP-based labeler.

133 E) Twitter COVID-19 CXR dataset: A collection of 135 CXRs showing COVID-19-related viral
134 pneumonia, collected from SARS-CoV-2 positive subjects has been made available by a
135 cardiothoracic radiologist from Spain via Twitter. (<https://twitter.com/ChestImaging>) The images are
136 made available in JFIF format at approximately 2K×2K resolution.

137 F) Montreal COVID-19 CXR dataset: As of April 14, 2020, a collection of 179 SARS-CoV-2
138 positive CXRs and others showing non-COVID-19 viral disease manifestations has been made
139 publicly available by the authors of [15] in their GitHub repository. The CXRs are made available in
140 AP and PA projections.

141 Table 1 shows the distribution of data extracted from the datasets identified above and used for
142 the different stages of learning performed in this study. The numerator and denominator show the
143 number of train and test data used in models' training and evaluations. The GT disease bounding
144 box annotations for a sample of the test data, containing 27 CXRs collectively from the Twitter
145 COVID-19 and Montreal COVID-19 CXR collections is set by the verification of publicly identified
146 cases from an expert radiologist who annotated the sample test collection.

147 **Table 1.** Dataset characteristics. Numerator and denominator denote the number of train and test
148 data respectively (UP=Pneumonia of unknown type, BP= Bacterial (proven) pneumonia, VP= non-
149 COVID-19 viral (proven) pneumonia, CP = COVID-19 pneumonia).

Dataset	UP	BP	VP	CP
A	-	2538/242	1345/148	-
B	-/6012	-	-	-
C	-/4683	-	-	-
D	-/307	-	-	-
E	-	-	-	-/135
F	-	-	-	-/179

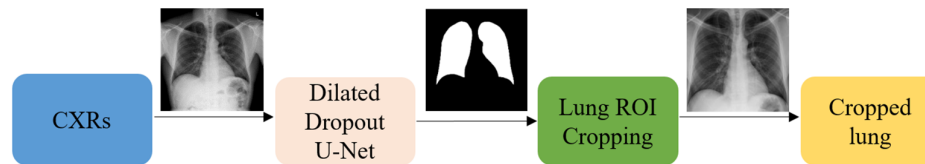
150 Broadly, our workflow consists of the following steps: First, we preprocess the images to make
151 them suitable for use in DL. Then, we evaluate the performance of a custom CNN and a selection of
152 pre-trained CNN models for binary categorization of the publicly available pediatric CXR collection
153 showing bacterial or viral pneumonia. The trained model is further used to categorize the publicly
154 available COVID-19 CXR collections as showing viral pneumonia. Next, we use the trained model to
155 weakly label the CXRs in the publicly available CXR collections with pneumonia-related opacities as
156 showing bacterial or viral pneumonia. The baseline training data is augmented with these weakly
157 labeled CXRs to improve detection performance with the baseline hold-out test data and the COVID-
158 19 CXR collections. We also augment the baseline training with COVID-19 CXRs from one of the two
159 different collections to evaluate for an improvement in performance in detecting CXRs showing
160 COVID-19 viral pneumonia from the other collection. This data augmentation strategy recognizes
161 the biological similarity in viral pneumonia and radiological manifestation due to COVID-19 caused
162 respiratory disease. It also takes advantage of dissimilarity to bacterial pneumonia-related opacities.
163 Finally, the strategy reduces the intra-class similarity and enhances inter-class discrimination in the
164 strategic ordering of the coarsely labeled data. We have already shown in our other work that
165 iteratively pruned deep learning ensembles produce impressive results with this data [6]. In this
166 work, we show that it is also possible to obtain very good results using a biologically sensitive and
167 discriminative training data augmentation strategy.
168

169 2.2. Lung ROI Segmentation and Preprocessing

170 It is important to add controls during training data-driven DL methods for disease
171 screening/diagnosis. Learning irrelevant feature representations could adversely impact clinical
172 decision making. To assist the DL model to focus on pulmonary abnormalities, we used a dilated

173 dropout-U-Net [16] to segment the lung ROI from the background. Dilated convolutions are shown
174 to improve performance [17] with exponential receptive field expansion while preserving spatial
175 resolution with no added computational complexity. A Gaussian dropout with an empirically
176 determined value of 0.2 is used after the convolutional layers in the network encoder to avoid
177 overfitting and improve generalization. A publicly available collection of CXRs and their associated
178 lung masks [18] is used to train the dilated dropout-U-Net model to generate lung masks of 224×224
179 pixel resolution. Callbacks are used to store the best model weights after each epoch. The generated
180 masks are superimposed on the original CXRs to delineate the lung boundaries, crop them to the size
181 of a bounding box, and re-scale them to 224×224 pixel resolution to reduce computational complexity.
182 Fig. 2 shows the segmentation steps performed in this study.

183 Additional preprocessing steps performed are as follows: i) CXRs are thresholded at to remove
184 very bright pixels to remove text annotations (empirically determined to be in the range [235 255])
185 that might be present in the cropped images. Missing pixels are in-painted using the surrounding
186 pixel values. ii) Images are normalized to make the pixel values lie in the range [0, 1]. iii) CXR images
187 are median filtered to remove noise and preserve edges. iv) Image pixel values are centered and
188 standardized to reduce computational complexity. Next, the cropped CXRs are used to train and
189 evaluate a custom CNN and a selection of pretrained models at different learning stages performed
190 in this study.
191



192
193

194 **Figure 2.** The segmentation approach showing dilated dropout U-Net based mask generation and
195 Lung ROI cropping.

196 2.3. Models and Computational Resources

197 The performance of a custom CNN model whose design is inspired by wide residual network
198 (WRN) architecture proposed in [19] and a selection of ImageNet pretrained CNN models is
199 evaluated during different stages of learning performed in this study. The benefit of using a WRN
200 compared to the traditional residual networks (ResNets) [20] is that it is shallower resulting in shorter
201 training times while producing similar or improved accuracy. In this study, we used a WRN based
202 custom CNN architecture with dropouts used in every residual block. After pilot empirical
203 evaluations, we used a network depth of 28, a width of 10, and a dropout ratio of 0.3 for the custom
204 WRN used in this study.

205 We evaluated the performance of the following pretrained CNN models, viz., a) VGG-16 [21], b)
206 Inception-V3 [22], c) Xception [23], d) DenseNet-121 [24], and e) NasNet-mobile [25]. The pretrained
207 CNNs are instantiated with their ImageNet [9] pretrained weights and truncated at their fully-
208 connected layers. The output feature maps are global average pooled and fed to a final dense layer
209 with Softmax activations to output the prediction probabilities.

210 The following hyperparameters of the custom WRN and pretrained CNNs are optimized
211 through a randomized grid search method: i) momentum, ii) L2-weight decay, and iii) initial learning
212 rate of the Stochastic Gradient Descent (SGD) optimizer. We initialized the search ranges to [0.80
213 0.99], [1e-8 1e-2], and [1e-7 1e-3] and for the learning momentum, L2-weight decay, and initial
214 learning rate, respectively. The custom WRN is initialized with random weights and the pretrained
215 models are fine-tuned end-to-end with smaller weight updates to make them data-specific and
216 classify the CXRs to their respective categories. Callbacks are used to monitor model performance
217 and store the best model weights for further analysis.

218 The performance of the custom WRN and the pretrained CNN models are evaluated in terms of
219 i) accuracy, ii) area under the (receiver operating characteristic -- ROC) curve (AUC), ii) sensitivity or

220 recall, iv) specificity, v) precision, vi) F-score, and vii) Mathews correlation coefficient (MCC). The
221 models are trained and evaluated on a Windows System with Intel Xeon CPU 3.80 GHz with 32 GB
222 RAM and NVIDIA GeForce GTX 1070 GPU. We used Keras 2.2.4 API version with Tensorflow
223 backend and CUDA/CUDNN dependencies.

224 *2.4. Weakly-labeled Data Augmentation*

225 We train the custom WRN and the pretrained models on the pediatric CXR collection [4] and
226 evaluated them on the ability to categorize hold-out test data into bacterial and viral pneumonia
227 categories. This start stems from following the literature which reveals that CXRs showing COVID-
228 19 viral pneumonia manifestations are visually similar to, yet distinct from those caused by bacterial,
229 fungal, and other non-COVID-19-related viral pneumonia [2]. We use the best performing baseline
230 model to evaluate its performance in categorizing the CXRs from Twitter COVID-19 and Montreal
231 COVID-19 collections as belonging to the viral pneumonia category.

232 We also evaluated the performance of the best performing baseline model in weakly
233 categorizing the CXRs showing pneumonia-related opacities from RSNA, CheXpert, and NIH CXR
234 collections as belonging to the bacterial or viral pneumonia categories. These weakly classified CXRs
235 are used to augment the baseline training data. The idea behind this augmentation is to expand the
236 training data feature space: i) to make the training distribution encompass the variability in the test
237 distribution, enhance inter-class discrimination, and reduce intra-class similarity; and, ii) to decrease
238 the generalization error by training with samples from a diversified distribution. The model is trained
239 with different combinations of the augmented training data and evaluated for an improvement in
240 performance as compared to the non-augmented baseline in classifying: i) the baseline hold-out
241 pediatric CXR test data to bacterial or viral pneumonia categories; and, ii) Twitter COVID-19 and
242 Montreal COVID-19 CXR collections as belonging to the viral pneumonia category. The baseline
243 training data is also augmented with the CXRs showing COVID-19 viral pneumonia from one of the
244 two different COVID-19 CXR collections used in this study to evaluate for performance improvement
245 with the other collection. This is done to evaluate if the COVID-19 viral pneumonia patterns are very
246 distinct and unique that can only improve performance toward COVID-19 detection as compared to
247 that with weakly-labeled data augmentation and non-augmented training.

248 *2.5. Salient ROI Localization*

249 Visualization helps in interpreting the model predictions and identify the salient ROI involved
250 in decision-making. In this study, the learned behavior of the best performing baseline model in
251 categorizing the CXRs to the bacterial and viral pneumonia classes is visualized through gradient-
252 weighted class activation maps (Grad-CAM) [26]. Grad-CAM is a gradient-based visualization
253 method where the gradients for a given class are computed concerning the features extracted from
254 the deepest convolutional layer in a trained model and are fed to a global average pooling layer to
255 obtain the weights of importance involved in decision-making. This results in a two-dimensional heat
256 map which is a weighted combination of the feature maps involved in categorizing the image to its
257 respective class.

258 **3. Results and Discussion**

259 Optimal hyperparameters values obtained using a randomized grid search for the custom WRN
260 and pretrained CNNs that are trained and evaluated on the pediatric CXR collection to classify them
261 at the patient level into showing bacterial or viral pneumonia are shown in Table 2. For model
262 validation, we allocated 20% of the training data which was randomly selected. The performance
263 achieved by the models is shown in Table 3.

264 It can be observed that the VGG-16 model demonstrates superior performance in terms of
265 accuracy and AUC with the hold-out test data. Xception model gives higher precision and specificity
266 than the other models. However, considering the F-score and MCC that give a balanced precision
267 and sensitivity measure, the VGG-16 model outperformed the others in classifying the pediatric CXRs

268 as showing bacterial or viral pneumonia. The performance excellence of the VGG-16 model is
 269 attributed to the fact that the architecture depth of the model is optimal to learn from the data used
 270 in this study and extract diversified features to categorize the CXRs to their respective categories.
 271 Deeper models like DenseNet-121 showed performance degradation as they suffered from overfitting
 272 issues and are not able to effectively model the variations across the categories. In this regard, we
 273 select the best performing VGG-16 model for further analysis on the Twitter COVID-19 and Montreal
 274 COVID-19 CXR collections as showing viral pneumonia.

275 **Table 2.** Optimal values for the hyperparameters for the custom WRN and pretrained CNNs
 276 obtained through randomized grid search M: Momentum, ILR: Initial learning rate, and L2: L2-
 277 weight decay).

Models	Optimal values		
	M	ILR	L2
Custom	0.90	1e-3	1e-5
Pretrained	0.95	1e-3	1e-6

278

279 **Table 3.** Performance achieved by the custom WRN and pretrained CNNs in classifying the
 280 pediatric CXR dataset into bacterial and viral categories. Here Acc.: Accuracy, Sens.: Sensitivity,
 281 Prec.: Precision, F: F-score, and MCC: Matthews Correlation Coefficient). Here bold values indicate
 282 superior performance.

Models	Acc.	AUC	Sens.	Spec.	Prec.	F	MCC
Custom WRN	0.8974	0.9534	0.9381	0.8311	0.9008	0.9191	0.7806
VGG-16	0.9308	0.9565	0.9711	0.8649	0.9216	0.9457	0.8527
Inception-V3	0.9103	0.937	0.9587	0.8311	0.9028	0.9299	0.8085
Xception	0.9282	0.954	0.9546	0.8852	0.9315	0.9429	0.8469
DenseNet-121	0.9026	0.9408	0.967	0.7973	0.8864	0.925	0.7931
NASNet-mobile	0.9282	0.9479	0.9753	0.8514	0.9148	0.944	0.8477

283

284 In this part of the study, we establish a baseline using the learned representations for the viral
 285 pneumonia category from the pediatric CXR collection for identifying COVID-19 viral pneumonia-
 286 related manifestations in the aforementioned COVID-19 CXR collections. As mentioned before, this
 287 is based on the knowledge that COVID-19 is a kind of viral pneumonia, but while being similar is
 288 different in some respects [2]. The baseline performance achieved is shown in Table 4. Fig. 3 shows
 289 the confusion matrix obtained toward classifying Twitter and Montreal COVID-19 CXR collections
 290 as showing viral pneumonia using baseline VGG-16 model trained to separate bacterial from viral
 291 pneumonia in CXR images.

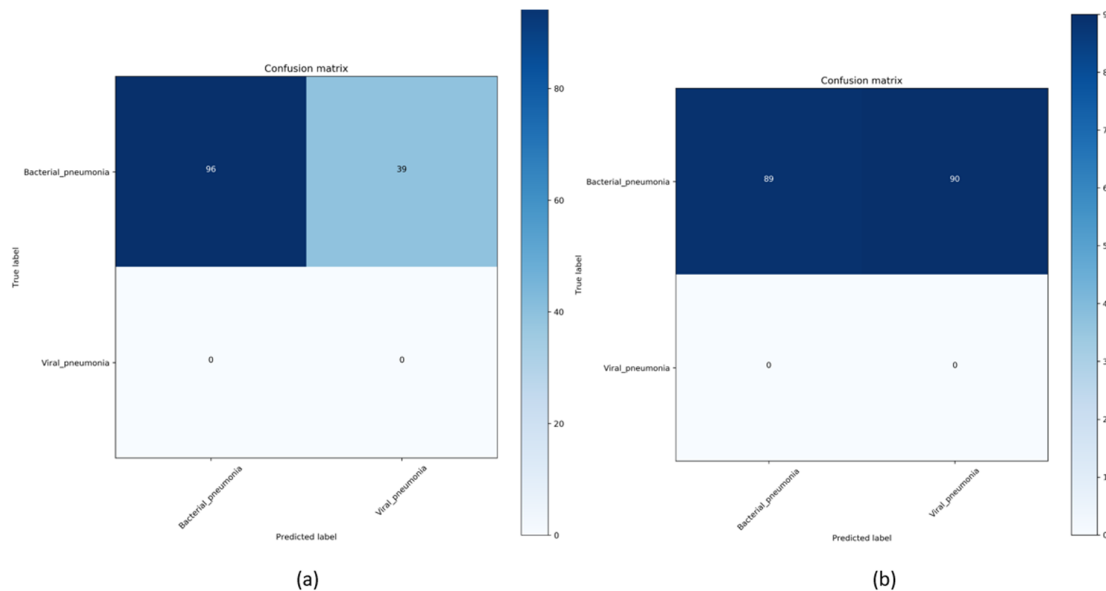
292 As observed in Table 4 and Fig. 3, the results obtained with the baseline VGG-16 model trained
 293 on the pediatric CXR collection to learn the representations of bacterial and viral pneumonia didn't
 294 deliver superior performance in detecting COVID-19 related viral pneumonia manifestations in the
 295 Twitter and Montreal COVID-19 CXR collections. We attribute this to limited variance in the training
 296 distribution and hence a narrow feature space to learn related patterns. The model fails to
 297 appropriately classify the Twitter and Montreal COVID-19 CXR collections as belonging to the viral
 298 pneumonia class.
 299

300
301
302

Table 4. Performance metrics achieved in classifying the Twitter and Montreal COVID-19 CXR collections as showing viral pneumonia using baseline VGG-16 model trained to separate bacterial from viral pneumonia in CXR images.

Model	Accuracy	
	Twitter-COVID-19	Montreal-COVID-19
VGG-16	0.2885	0.5028

303
304



305
306
307
308

Figure 3. Confusion matrix obtained toward classifying (a) Twitter and (b) Montreal COVID-19 CXR collections as showing viral pneumonia using baseline VGG-16 model.

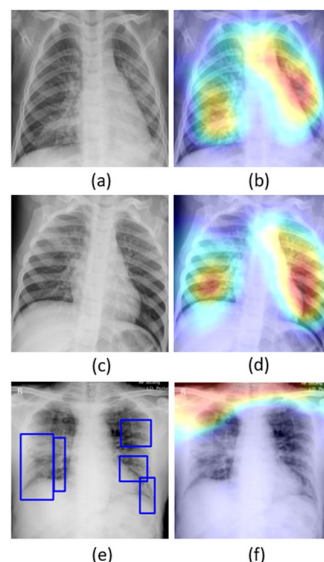
309
310
311
312
313
314
315
316
317
318
319
320
321

The learned behavior of the baseline trained VGG-16 model with the pediatric CXR collection is interpreted through Grad-CAM visualizations and is shown in Fig. 4. The gradients for the bacterial and viral pneumonia classes that are flowing into the deepest convolutional layer of the trained model are used to interpret the neurons involved in decision-making. The heat maps obtained as a result of weighing these feature maps are superimposed on the original CXRs to identify the salient ROI involved in categorizing the CXRs to their respective classes. It is observed that the model is correctly focusing on the salient ROI for the test data coming from the same training distribution that helps to categorize them into bacterial and viral pneumonia classes. However, the salient ROI involved in categorizing a test image from the Montreal COVID-19 CXR collection that comes from a different distribution compared to the baseline training data didn't properly overlap with the GT annotations. This leads to the inference that the model is not properly trained to identify the disease manifestations in the unseen test data that has similar, yet distinct visual representations as to the baseline training data.

322
323
324
325
326
327
328
329
330

With data-driven DL methods, the training data may contain samples that do not contribute to decision-making. Modifying the training distribution could provide an active solution to improve performance with similar and/or different test distribution. In response, our approach is to expand the training data feature space to create a diversified distribution that could help learn and improve the performance with the baseline test data coming from the same distribution as the training data and/or with other test data coming from a different distribution. In this study, we propose to expand the training data feature space by augmenting them with weakly classified CXR images. For this, the trained baseline VGG-16 model is used to weakly classify the CXR images from NIH, RSNA, and CheXpert collections showing pneumonia-related opacities as showing bacterial or viral pneumonia.

331 The weakly labeled images are further stored to augment the baseline training data to improve
 332 performance in categorizing the test data from pediatric, Twitter COVID-19, and Montreal COVID-
 333 19 CXR collections. We also augmented the baseline with the COVID-19 CXR collections to study
 334 their effect on improving performance with the baseline test data. The performance metrics achieved
 335 with the baseline test data using different combinations of the augmented training data is shown in
 336 Table 5.
 337



338 **Figure 4.** Original CXRs and their salient ROI visualization: (a) and (b) shows a CXR with bilateral
 339 bacterial pneumonia and the corresponding Grad-CAM visualization; (c) and (d) shows a CXR with
 340 viral pneumonia manifestations and the corresponding salient ROI visualization; (e) and (f) shows a
 341 sample CXR from the test set of Montreal COVID-19 CXR collection with GT annotations and the
 342 corresponding salient ROI visualization.
 343

344 **Table 5.** Performance metrics achieved with the different combinations of the augmented training
 345 data toward classifying the baseline test data into bacterial and viral pneumonia categories. Bold
 346 values indicate superior performance.

Dataset	Acc.	AUC	Sens.	Spec.	Prec.	F	MCC
Baseline	0.9308	0.9565	0.9711	0.8649	0.9216	0.9457	0.8527
Data augmentation with weakly labeled images							
Baseline + Montreal	0.9179	0.9479	0.9794	0.8176	0.8978	0.9368	0.827
Baseline + Twitter	0.9308	0.9577	0.9835	0.8446	0.9119	0.9464	0.8541
Baseline + NIH	0.9179	0.9600	0.9587	0.8514	0.9134	0.9355	0.8249
Baseline + CheXpert	0.9405	0.9689	0.9877	0.8624	0.9201	0.9542	0.8716
Baseline + RSNA	0.9359	0.9592	0.9877	0.8514	0.9158	0.9503	0.8653
Baseline + NIH + CheXpert	0.9333	0.9606	0.9835	0.8514	0.9154	0.9483	0.8594
Baseline + NIH + RSNA	0.9231	0.9642	0.9959	0.8041	0.8926	0.9415	0.8411
Baseline + CheXpert + RSNA	0.9359	0.9628	0.9835	0.8582	0.919	0.9501	0.8647
Baseline + NIH + CheXpert + RSNA	0.9154	0.9542	0.9794	0.8109	0.8944	0.935	0.8217
Baseline + CheXpert + Twitter	0.9103	0.9538	0.9629	0.8244	0.8997	0.9302	0.8088
Baseline + CheXpert + Montreal	0.9231	0.9595	0.9711	0.8446	0.9109	0.94	0.8365

347

348 Note that the baseline training data augmented with the weakly labeled CXR images from the
 349 CheXpert CXR collection demonstrated superior performance in all metrics compared to the non-
 350 augmented and other combinations of augmented training data. This underscores the fact that this
 351 augmentation approach resulted in a favorable increase in the training data size, encompassing a
 352 diversified distribution to learn and improve performance in the test data, compared to that with
 353 non-augmented training.

354 We also studied the effect of weakly labeled data augmentation with the test data from Twitter
 355 and Montreal COVID-19 CXR collections. The results are as shown in Table 6.

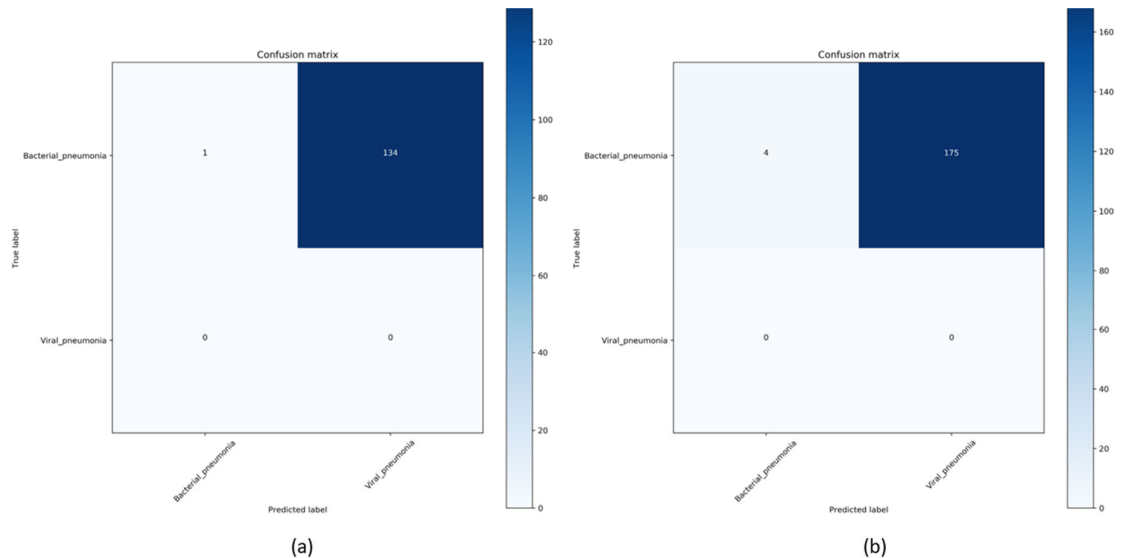
356 **Table 6.** Performance metrics achieved using combinations of the augmented training data toward
 357 classifying Twitter and Montreal COVID-19 CXR collections as belonging to the viral pneumonia
 358 category. Bold values indicate superior performance.

Dataset	Accuracy	
	Twitter-COVID-19	Montreal-COVID-19
Baseline	0.2885	0.5028
Data augmentation with weakly labeled images		
Baseline + NIH	0.1037	0.2625
Baseline + CheXpert	0.5555	0.6536
Baseline + RSNA	0.2296	0.4469
Baseline + NIH + CheXpert	0.1852	0.4078
Baseline + NIH + RSNA	0.1407	0.4413
Baseline + CheXpert + RSNA	0.2222	0.4357
Baseline + NIH + CheXpert + RSNA	0.1852	0.4413
Baseline + CheXpert + Twitter	-	0.7095
Baseline + CheXpert + Montreal	0.8889	-
Baseline + Twitter	-	0.9778
Baseline + Montreal	0.9926	-

359 The performance evaluation results demonstrate that the baseline training data augmented with the
 360 weakly labeled CXR images from the CheXpert collection initially improved performance with an
 361 accuracy of 0.5555 and 0.6536 as compared to the non-augmented baseline (0.2885 and 0.5028) in
 362 classifying Twitter and Montreal COVID-19 CXR test data, respectively, as belonging to the viral
 363 pneumonia category. The performance degradation with other combinations of weakly-labeled data
 364 augmentation underscores the fact that adding more data introduces noise into the training process;
 365 increasing the number of training samples do not always improve performance since these samples
 366 either do not contribute or adversely impact decision-making.

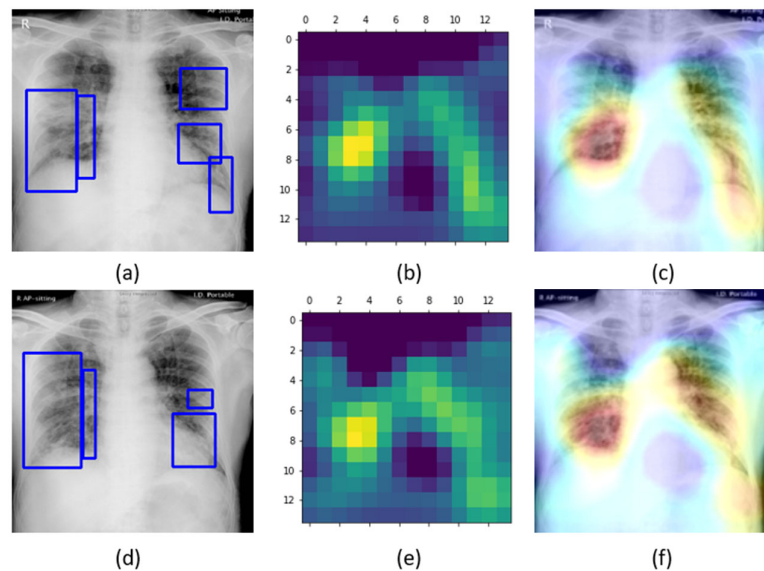
368 Modifying the distribution of the training data in a way to include only those samples could
 369 provide an effective solution to improve performance with the test data from a similar or different
 370 distribution as compared to the non-augmented training data. In this regard, we also augmented the
 371 baseline training data with the COVID-19 viral pneumonia CXRs from one of two different collections
 372 and evaluated the performance with the other. This is performed to evaluate if there is a performance
 373 improvement if the training data is modified to include only samples with a known, similar
 374 distribution. It is observed from Table 6 that augmenting the baseline training data with the Twitter
 375 COVID-19 CXR collection significantly improved performance in detecting COVID-19 CXRs in the
 376 Montreal collection as compared to the weakly-labeled augmentation using CheXpert CXRs and the
 377 non-augmented baseline. We observed similar improvements in performance with the Twitter
 378 COVID-19 CXRs when the baseline training data is augmented with the Montreal COVID-19 CXR

379 collection for model training. Fig. 5 shows the confusion matrix obtained toward this study. This
380 underscores the fact that augmenting the training data with COVID-19 CXRs, though not coming
381 from the same collection, significantly improved performance with the test data from a different
382 COVID-19 CXR collection, as compared to non-augmented baseline and weakly-labeled data
383 augmentation with non-COVID-19 viral and bacterial pneumonia CXRs. The COVID-19 viral
384 pneumonia has a distinct pattern, compared to non-COVID-19 viral and other pneumonia. For this
385 reason, irrespective of the collection the CXRs come from, augmenting the training data with samples
386 from one COVID-19 CXR collection significantly improves performance with the other.
387



388
389 **Figure 5.** Confusion matrix obtained toward classifying (a) Montreal COVID-19 CXR
390 collections as showing viral pneumonia using the VGG-16 model trained on the baseline augmented
391 with Montreal COVID-19 and Twitter COVID-19 CXR collections, respectively.
392

393 Fig.6 shows the learned behavior of the VGG-16 model trained on the baseline data augmented
394 with Montreal COVID-19 and Twitter COVID-19 CXR collections individually to predict on a test
395 sample with GT annotations from Montreal COVID-19 and Twitter COVID-19 CXR collections,
396 respectively.
397



398

399 **Figure 6.** Original CXRs, heat maps, and salient ROI visualization: (a), (b), and (c) shows a COVID-
400 19 viral pneumonia test CXR from Montreal collection with GT annotations, the corresponding heat
401 map, and Grad-CAM visualization, (d), (e), and (f) shows a COVID-19 viral pneumonia test CXR
402 from the Twitter collection with GT annotations, the heat map, and its associated class activation
403 maps.
404

405 Unlike the degraded performance of the model trained on non-augmented data that failed to
406 localize salient ROI in a test CXR showing COVID-19 viral pneumonia, as observed from Fig. 4, the
407 model trained on the augmented baseline with COVID-19 CXRs from one collection delivered
408 superior localization performance with the test CXR samples from the other collection. Fig. 6a shows
409 the learned interpretation of these trained models in the form of heat maps and class activation maps.
410 It is observed that the models are correctly focusing on the salient ROI, matching with the GT
411 annotations that help to categorize them as showing COVID-19 viral pneumonia. This leads to the
412 inference that the model has effectively learned the diversified feature space augmented with class-
413 specific (COVID-19 viral pneumonia) data that has a distinct pattern compared to non-COVID-19
414 viral and bacterial pneumonia to effectively localize the salient ROI involved in decision-making.

415 **4. Conclusions and Future Work**

416 Image Weakly labeled data augmentation helped to improve performance with the hold-out
417 baseline test data because the CXRs with pneumonia-related opacities in CheXpert collection has a
418 similar distribution to bacterial and non-COVID-19 viral pneumonia that helped to expand the
419 training feature space by introducing a controlled variance to improve performance with the baseline
420 test data. However, with COVID-19 CXRs, weakly-labeled data augmentation didn't deliver superior
421 performance since COVID-19 viral pneumonia has a distinct pattern as compared to non-COVID-19
422 viral and bacterial pneumonia.

423 In this study, we evaluated the effect of weakly-labeled data augmentation toward classifying
424 the CXRs as showing COVID-19 viral pneumonia. In this regard, being a one-class problem, we have
425 only false-negatives and no false positives. As future work, we aim to expand the analysis toward
426 classifying non-COVID-19 and COVID-19 viral pneumonia and other multi-class problems, where
427 we aim to perform multi-class ROC analysis and obtain an efficient operating point suiting model
428 deployment. Considering limited data availability as with COVID-19 detection, we also aim to
429 construct model ensembles to combine the predictions of models trained on various combinations of
430 augmented training data to further improve COVID-19 detection performance.

431 **Author Contributions:** Conceptualization, Sivaramakrishnan Rajaraman; Data curation, Sivaramakrishnan
432 Rajaraman; Formal analysis, Sivaramakrishnan Rajaraman; Funding acquisition, Sameer Antani; Investigation,
433 Sivaramakrishnan Rajaraman and Sameer Antani; Methodology, Sivaramakrishnan Rajaraman and Sameer
434 Antani; Project administration, Sameer Antani; Resources, Sameer Antani; Software, Sivaramakrishnan
435 Rajaraman; Supervision, Sivaramakrishnan Rajaraman and Sameer Antani; Visualization, Sivaramakrishnan
436 Rajaraman; Writing – original draft, Sivaramakrishnan Rajaraman; Writing – review & editing,
437 Sivaramakrishnan Rajaraman and Sameer Antani. All authors have read and agreed to the published version of
438 the manuscript.

439 **Funding:** This work was supported by the Intramural Research Program of the Lister Hill National Center for
440 Biomedical Communications (LHNCBC), the National Library of Medicine (NLM), and the U.S. National
441 Institutes of Health (NIH).

442 **Acknowledgments:** We are grateful to Dr. Jenifer Siegelman of Takeda Pharmaceuticals for her radiological
443 expertise in annotating a sample of COVID-19 test data and discussions related to the radiology of COVID-19.

444 **Conflicts of Interest:** The authors declare no conflict of interest.

445
446
447

448 References

- 449 1. World Health Organization (WHO). Coronavirus disease (COVID-2019) situation reports.
450 <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (archived on 2nd
451 March 2020).
- 452 2. Rubin, G.D. et al. The Role of Chest Imaging in Patient Management during the COVID-19 Pandemic: A
453 Multinational Consensus Statement from the Fleischner Society. *Radiology*, **2020**, 201365.
- 454 3. Bai, Harrison X. et al., Performance of Radiologists in Differentiating COVID-19 from Viral Pneumonia On
455 Chest CT. *Radiology*, **2020**, 200823.
- 456 4. Kermany, S. et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.
457 *Cell*, **2018**, 172(5), 1122-1131.
- 458 5. Maghddid, H.S. et al. Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning
459 and Transfer Learning Algorithms. *arXiv Preprint arXiv:2004.00038*, **2020**.
- 460 6. Rajaraman, S.; Siegelman, J.; Alderson, P.O.; Folio, L.S.; Folio, .R.; Antani, S.K. Iteratively Pruned Deep
461 Learning Ensembles for COVID-19 Detection in Chest X-rays. *arXiv Preprint arXiv:2004.08379*, **2020**.
- 462 7. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv*
463 *Preprint arXiv:1712.04621*, **2017**.
- 464 8. Lv, X.; Zhang, X. Generating Chinese Classical Landscape Paintings Based on Cycle-Consistent
465 Adversarial Networks. *Proc. 6th International Conference on Systems and Informatics (ICSAI)*, **2019**, 1265-1269.
- 466 9. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database.
467 *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **2009**, 248-255.
- 468 10. Cohen, A.-B.; Klang, E.; Amitai, M.M.; Goldberger, J.; Greenspan, H. Anatomical data augmentation for
469 CNN based pixel-wise classification. *Proc. IEEE 15th International Symposium on Biomedical Imaging (ISBI)*,
470 **2018**, 1096-1099.
- 471 11. Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv Preprint arXiv:1701.00160*, **2016**.
- 472 12. Shih, G. et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert
473 Annotations of Possible Pneumonia. *Radiol Artif Intell.*, **2019**, 1 (1), 1-5.
- 474 13. Wang, X. et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised
475 Classification and Localization of Common Thorax Diseases. *Proc. Int. Conf. Computer Vision (ICCV)*, **2017**,
476 3462-3471.
- 477 14. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Illcus, S.; Chute, C.; Marklund, H.; Haghighi, B.; Ball, R.;
478 Shpanskaya, K.; Seekins, J.; Mong, D.A.; Halabi, S.S.; Sandberg, J.K.; Jones, R.; Larson, D.B.; Langlotz, C.P.;
479 Patel, B.N.; Lungren, M.P.; Ng, A.Y. CheXpert: a large chest radiograph dataset with uncertainty labels and
480 expert comparison. *Proc. Thirty-third AAAI conference on artificial intelligence (AAAI)*, **2019**, 590-597.
- 481 15. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv Preprint arXiv:2003.11597*, **2020**.
- 482 16. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: convolutional networks for biomedical image segmentation.
483 *arXiv Preprint arXiv:1505.04597*, **2015**.
- 484 17. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv Preprint*
485 *arXiv:1511.07122*, **2015**.
- 486 18. Candemir, S.; Antani, S.K.; Jaeger, S.R.; Thoma, G.R. Lung boundary detection in pediatric chest X-rays.
487 *Proc. SPIE. Medical imaging: PACS and imaging informatics: next generation and innovations*, **2015**, 94180Q.
- 488 19. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. *arXiv Preprint arXiv:1605.07146*, **2017**.
- 489 20. Zhang H.X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proc. Int. Conf. Computer Vision*
490 *(ICCV)*, **2016**, 770-778.
- 491 21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *Proc. Int.*
492 *Conf. Learning Representations (ICLR)*, **2015**, 1-14.
- 493 22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception architecture for
494 computer vision. *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, **2016**, 2818-2826.
- 495 23. Chollet, F. Xception: Deep learning with depthwise separable convolutions. *IEEE Conf. Computer Vision and*
496 *Pattern Recognition (CVPR)*, **2017**, 1251-1258.
- 497 24. Liu, H.Z.; van der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. *Proc. Int. Conf.*
498 *Computer Vision (ICCV)*, **2017**, 4700-4708.
- 499 25. Pham, M.Y.; Zoph, G.B.; Le, Q.V.; Dean, J. Efficient neural architecture search via parameter sharing. *Proc.*
500 *Int. Conf. Machine Learning (ICML)*, **2018**, 4092-4101.
- 501 26. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations
502 from deep networks via gradient-based localization. *Proc. Int. Conf. Computer Vision (ICCV)*, **2017**, 618-626.